Coursera Capstone Project Report
IBM Applied Data Science

Finding ideal House in New Delhi

By: Siddhant Bhanot

May 2020

Introduction:

When looking to buy a house, location of the house and the cost are two of the major factors to be considered.

Location plays an important role as one wants a house a location having the maximum amenities along with a suitable price value.

In this project I have classified various localities of the city of New Delhi, India based on the amenities offered and the average cost of the houses.

Business Problem:

The objective of this capstone project is to analyse and select the best location suitable for a person to buy a house, based on the amenities offered in that location and price of the house.

Target audience:

This project will be useful for anybody looking to buy a house in New Delhi.

Data:

The following data is required

- *List of neighbourhoods along with average price per square feet of the house for the neighbourhood for the city of New Delhi*

- *Latitude and Longitude of the given localities*

- *Amenities around the given location.*

Sources of Data:

From the website, https://www.makaan.com/price-trends/property-rates-for-buy-in-delhi?page=1, I have got the list of neighbourhoods along with their average price per square feet of the houses in those neighbourhood .

Then, in order to get the geographical coordinates of the given neighbourhoods, I have used python Geocoder package.

Finally, in order to explore the given neighbourhoods, I have used FourSquare api.

Methodology:

Using BeautifulSoup, I have scraped the site https://www.makaan.com/price-trends/property-rates-for-buy-in-delhi?page=1, for the data on localities and average price of the houses.

After cleaning the scraped data, I included the latitudes and longitudes of the localities and applied FourSquare api, to get the top 100 venues within a radius of 500 m.

Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering.

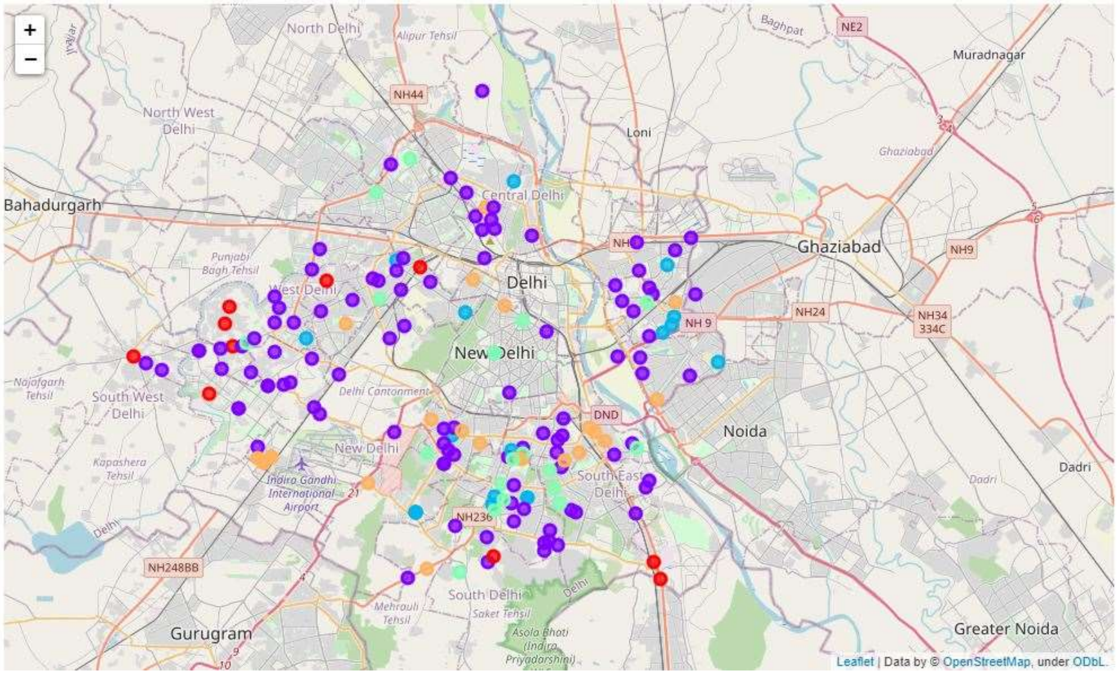
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 5 clusters.

Result:

The results from the k-means clustering show that we can categorize the neighbourhoods into 5 clusters based on the frequency of occurrence of different venues and the average price of houses in that neighbourhood.

- Cluster 1 is the most cheap
- Cluster 2,3 and 4 are have moderate prices of houses.
- Cluster 5 has the highest prices of houses.

The results of the clustering are visualized in the map below with cluster 0 in red colour, 1 in purple, 2 in blue, 3 in green and 4 in orange


+
−
North Delhi
Alipur Tehsil
Baghpat
NE2
Muradnagar
NH44
North West Delhi
Loni
Ghaziabad
Bahadurgarh
Central Delhi
NH9
Ghaziabad
NH9
Punjabi Bagh Tehsil
West Delhi
Delhi
NH24
NH34 334C
NH 9
New Delhi
Najafgarh Tehsil
South West Delhi
Delhi Cantonment
DND
Noida
Kapashera Tehsil
New Delhi
Indira Gandhi International Airport
South East Delhi
Dadri
Dadri
NH236
Delhi
NH248BB
South Delhi
Saket Tehsil
Mehrauli Tehsil
Delhi
Asola Bhati (Indira Priyadarshini)
Gurugram
Greater Noida
Leaflet | Data by © OpenStreetMap, under ODbL.